IBM SPSS Complex Samples 22

Note
Before using this information and the product it supports, read the information in "Notices" on page 51.

Product Information

This edition applies to version 22, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Introduction to Complex Samples Procedures

An inherent assumption of analytical procedures in traditional software packages is that the observations in a data file represent a simple random sample from the population of interest. This assumption is untenable for an increasing number of companies and researchers who find it both cost-effective and convenient to obtain samples in a more structured way.

The Complex Samples option allows you to select a sample according to a complex design and incorporate the design specifications into the data analysis, thus ensuring that your results are valid.

Properties of Complex Samples

A complex sample can differ from a simple random sample in many ways. In a simple random sample, individual sampling units are selected at random with equal probability and without replacement (WOR) directly from the entire population. By contrast, a given complex sample can have some or all of the following features:

Stratification. Stratified sampling involves selecting samples independently within non-overlapping subgroups of the population, or strata. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups. With stratification, you can ensure adequate sample sizes for subgroups of interest, improve the precision of overall estimates, and use different sampling methods from stratum to stratum.

Clustering. Cluster sampling involves the selection of groups of sampling units, or clusters. For example, clusters may be schools, hospitals, or geographical areas, and sampling units may be students, patients, or citizens. Clustering is common in multistage designs and area (geographic) samples.

Multiple stages. In multistage sampling, you select a first-stage sample based on clusters. Then you create a second-stage sample by drawing subsamples from the selected clusters. If the second-stage sample is based on subclusters, you can then add a third stage to the sample. For example, in the first stage of a survey, a sample of cities could be drawn. Then, from the selected cities, households could be sampled. Finally, from the selected households, individuals could be polled. The Sampling and Analysis Preparation wizards allow you to specify three stages in a design.

Nonrandom sampling. When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

Unequal selection probabilities. When sampling clusters that contain unequal numbers of units, you can use probability-proportional-to-size (PPS) sampling to make a cluster's selection probability equal to the proportion of units it contains. PPS sampling can also use more general weighting schemes to select units.

Unrestricted sampling. Unrestricted sampling selects units with replacement (WR). Thus, an individual unit can be selected for the sample more than once.

Sampling weights. Sampling weights are automatically computed while drawing a complex sample and ideally correspond to the "frequency" that each sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size. Complex Samples analysis procedures require sampling weights in order to properly analyze a complex sample. Note that these weights should be used entirely within the Complex Samples option and should not be used with other analytical procedures via the Weight Cases procedure, which treats weights as case replications.

Usage of Complex Samples Procedures

Your usage of Complex Samples procedures depends on your particular needs. The primary types of users are those who:

- Plan and carry out surveys according to complex designs, possibly analyzing the sample later. The primary tool for surveyors is the Sampling Wizard.
- Analyze sample data files previously obtained according to complex designs. Before using the Complex Samples analysis procedures, you may need to use the Analysis Preparation Wizard.

Regardless of which type of user you are, you need to supply design information to Complex Samples procedures. This information is stored in a **plan file** for easy reuse.

Plan Files

A plan file contains complex sample specifications. There are two types of plan files:

Sampling plan. The specifications given in the Sampling Wizard define a sample design that is used to draw a complex sample. The sampling plan file contains those specifications. The sampling plan file also contains a default analysis plan that uses estimation methods suitable for the specified sample design.

Analysis plan. This plan file contains information needed by Complex Samples analysis procedures to properly compute variance estimates for a complex sample. The plan includes the sample structure, estimation methods for each stage, and references to required variables, such as sample weights. The Analysis Preparation Wizard allows you to create and edit analysis plans.

There are several advantages to saving your specifications in a plan file, including:

- A surveyor can specify the first stage of a multistage sampling plan and draw first-stage units now, collect information on sampling units for the second stage, and then modify the sampling plan to include the second stage.
- An analyst who doesn't have access to the sampling plan file can specify an analysis plan and refer to that plan from each Complex Samples analysis procedure.
- A designer of large-scale public use samples can publish the sampling plan file, which simplifies the instructions for analysts and avoids the need for each analyst to specify his or her own analysis plans.

Further Readings

For more information on sampling techniques, see the following texts:

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. New York: John Wiley and Sons.

Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

Särndal, C., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Chapter 2. Sampling from a Complex Design

The Sampling Wizard guides you through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, you should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Creating a New Sample Plan

1. From the menus choose:

   Analyze > **Complex Samples** > **Select a Sample...**
2. Select **Design a sample** and choose a plan filename to save the sample plan.
3. Click **Next** to continue through the Wizard.
4. Optionally, in the Design Variables step, you can define strata, clusters, and input sample weights. After you define these, click **Next**.
5. Optionally, in the Sampling Method step, you can choose a method for selecting items.

   If you select **PPS Brewer** or **PPS Murthy**, you can click **Finish** to draw the sample. Otherwise, click **Next** and then:
6. In the Sample Size step, specify the number or proportion of units to sample.
7. You can now click **Finish** to draw the sample.

Optionally, in further steps you can:

- Choose output variables to save.
- Add a second or third stage to the design.
- Set various selection options, including which stages to draw samples from, the random number seed, and whether to treat user-missing values as valid values of design variables.
- Choose where to save output data.
- Paste your selections as command syntax.

Sampling Wizard: Design Variables

This step allows you to select stratification and clustering variables and to define input sample weights. You can also specify a label for the stage.

Stratify By. The cross-classification of stratification variables defines distinct subpopulations, or strata. Separate samples are obtained for each stratum. To improve the precision of your estimates, units within strata should be as homogeneous as possible for the characteristics of interest.

Clusters. Cluster variables define groups of observational units, or clusters. Clusters are useful when directly sampling observational units from the population is expensive or impossible; instead, you can sample clusters from the population and then sample observational units from the selected clusters. However, the use of clusters can introduce correlations among sampling units, resulting in a loss of precision. To minimize this effect, units within clusters should be as heterogeneous as possible for the characteristics of interest. You must define at least one cluster variable in order to plan a multistage design. Clusters are also necessary in the use of several different sampling methods. See the topic "Sampling Wizard: Sampling Method" on page 4 for more information.

Input Sample Weight. If the current sample design is part of a larger sample design, you may have sample weights from a previous stage of the larger design. You can specify a numeric variable containing these weights in the first stage of the current design. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. You can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

Note: The source variable list has the same content across steps of the Wizard. In other words, variables removed from the source list in a particular step are removed from the list in all steps. Variables returned to the source list appear in the list in all steps.

Tree Controls for Navigating the Sampling Wizard

On the left side of each step in the Sampling Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, if each previous step has been given the minimum required specifications for that step. See the Help for individual steps for more information on why a given step may be invalid.

Sampling Wizard: Sampling Method

This step allows you to specify how to select cases from the active dataset.

Method. Controls in this group are used to choose a selection method. Some sampling types allow you to choose whether to sample with replacement (WR) or without replacement (WOR). See the type descriptions for more information. Note that some probability-proportional-to-size (PPS) types are available only when clusters have been defined and that all PPS types are available only in the first stage of a design. Moreover, WR methods are available only in the last stage of a design.

- **Simple Random Sampling.** Units are selected with equal probability. They can be selected with or without replacement.
- **Simple Systematic.** Units are selected at a fixed interval throughout the sampling frame (or strata, if they have been specified) and extracted without replacement. A randomly selected unit within the first interval is chosen as the starting point.
- **Simple Sequential.** Units are selected sequentially with equal probability and without replacement.
- **PPS.** This is a first-stage method that selects units at random with probability proportional to size. Any units can be selected with replacement; only clusters can be sampled without replacement.
- **PPS Systematic.** This is a first-stage method that systematically selects units with probability proportional to size. They are selected without replacement.
- **PPS Sequential.** This is a first-stage method that sequentially selects units with probability proportional to cluster size and without replacement.
- **PPS Brewer.** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Murthy.** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Sampford.** This is a first-stage method that selects more than two clusters from each stratum with probability proportional to cluster size and without replacement. It is an extension of Brewer's method. A cluster variable must be specified to use this method.
- **Use WR estimation for analysis.** By default, an estimation method is specified in the plan file that is consistent with the selected sampling method. This allows you to use with-replacement estimation even if the sampling method implies WOR estimation. This option is available only in stage 1.

Measure of Size (MOS). If a PPS method is selected, you must specify a measure of size that defines the size of each unit. These sizes can be explicitly defined in a variable or they can be computed from the data. Optionally, you can set lower and upper bounds on the MOS, overriding any values found in the MOS variable or computed from the data. These options are available only in stage 1.

Sampling Wizard: Sample Size

This step allows you to specify the number or proportion of units to sample within the current stage. The sample size can be fixed or it can vary across strata. For the purpose of specifying sample size, clusters chosen in previous stages can be used to define strata.

Units. You can specify an exact sample size or a proportion of units to sample.

- **Value.** A single value is applied to all strata. If **Counts** is selected as the unit metric, you should enter a positive integer. If **Proportions** is selected, you should enter a non-negative value. Unless sampling with replacement, proportion values should also be no greater than 1.
- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

If **Proportions** is selected, you have the option to set lower and upper bounds on the number of units sampled.

Define Unequal Sizes

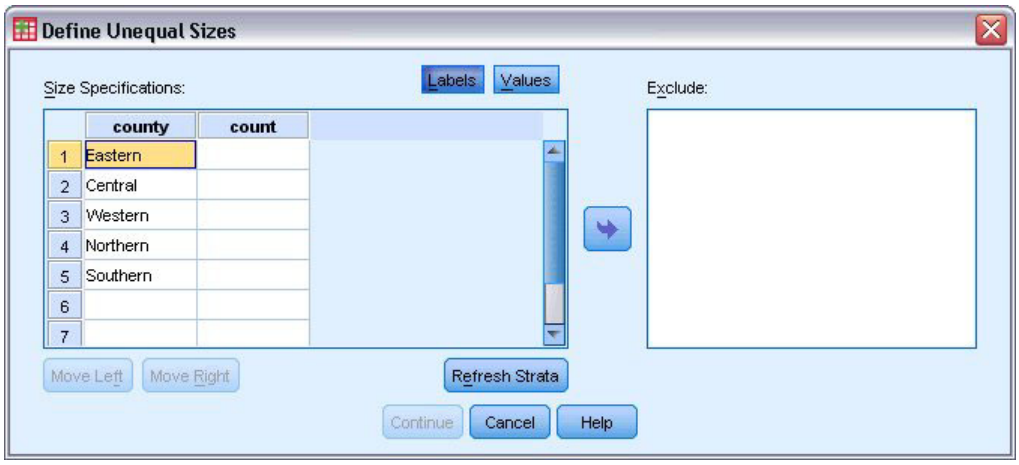

Figure 1. Define Unequal Sizes dialog box

The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

Size Specifications grid. The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click **Labels** or **Values** to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click **Refresh Strata** to repopulate the grid with each combination of labeled data values for variables in the grid.

Exclude. To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the Exclude list. These variables are not used to define sample sizes.

Sampling Wizard: Output Variables

This step allows you to choose variables to save when the sample is drawn.

Population size. The estimated number of units in the population for a given stage. The rootname for the saved variable is *PopulationSize_*.

Sample proportion. The sampling rate at a given stage. The rootname for the saved variable is *SamplingRate_*.

Sample size. The number of units drawn at a given stage. The rootname for the saved variable is *SampleSize_*.

Sample weight. The inverse of the inclusion probabilities. The rootname for the saved variable is *SampleWeight_*.

Some stagewise variables are generated automatically. These include:

Inclusion probabilities. The proportion of units drawn at a given stage. The rootname for the saved variable is *InclusionProbability_*.

Cumulative weight. The cumulative sample weight over stages previous to and including the current one. The rootname for the saved variable is *SampleWeightCumulative_*.

Index. Identifies units selected multiple times within a given stage. The rootname for the saved variable is *Index_*.

Note: Saved variable rootnames include an integer suffix that reflects the stage number—for example, *PopulationSize_1_* for the saved population size for stage 1.

Sampling Wizard: Plan Summary

This is the last step within each stage, providing a summary of the sample design specifications through the current stage. From here, you can either proceed to the next stage (creating it, if necessary) or set options for drawing the sample.

Sampling Wizard: Draw Sample Selection Options

This step allows you to choose whether to draw a sample. You can also control other sampling options, such as the random seed and missing-value handling.

Draw sample. In addition to choosing whether to draw a sample, you can also choose to execute part of the sampling design. Stages must be drawn in order—that is, stage 2 cannot be drawn unless stage 1 is also drawn. When editing or executing a plan, you cannot resample locked stages.

Seed. This allows you to choose a seed value for random number generation.

Include user-missing values. This determines whether user-missing values are valid. If so, user-missing values are treated as a separate category.

Data already sorted. If your sample frame is presorted by the values of the stratification variables, this option allows you to speed the selection process.

Sampling Wizard: Draw Sample Output Files

This step allows you to choose where to direct sampled cases, weight variables, joint probabilities, and case selection rules.

Sample data. These options let you determine where sample output is written. It can be added to the active dataset, written to a new dataset, or saved to an external IBM® SPSS® Statistics data file. Datasets

are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to variable naming rules. If an external file or new dataset is specified, the sampling output variables and variables in the active dataset for the selected cases are written.

Joint probabilities. These options let you determine where joint probabilities are written. They are saved to an external IBM SPSS Statistics data file. Joint probabilities are produced if the PPS WOR, PPS Brewer, PPS Sampford, or PPS Murthy method is selected and WR estimation is not specified.

Case selection rules. If you are constructing your sample one stage at a time, you may want to save the case selection rules to a text file. They are useful for constructing the subframe for subsequent stages.

Sampling Wizard: Finish

This is the final step. You can save the plan file and draw the sample now or paste your selections into a syntax window.

When making changes to stages in the existing plan file, you can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If you want to save the plan to a new file, select **Paste the syntax generated by the Wizard into a syntax window** and change the filename in the syntax commands.

Modifying an Existing Sample Plan

1. From the menus choose:

   Analyze > **Complex Samples** > **Select a Sample...**
2. Select **Edit a sample design** and choose a plan file to edit.
3. Click **Next** to continue through the Wizard.
4. Review the sampling plan in the Plan Summary step, and then click **Next**.

   Subsequent steps are largely the same as for a new design. See the Help for individual steps for more information.
5. Navigate to the Finish step, and specify a new name for the edited plan file or choose to overwrite the existing plan file.

Optionally, you can:

- Specify stages that have already been sampled.
- Remove stages from the plan.

Sampling Wizard: Plan Summary

This step allows you to review the sampling plan and indicate stages that have already been sampled. If editing a plan, you can also remove stages from the plan.

Previously sampled stages. If an extended sampling frame is not available, you will have to execute a multistage sampling design one stage at a time. Select which stages have already been sampled from the drop-down list. Any stages that have been executed are locked; they are not available in the Draw Sample Selection Options step, and they cannot be altered when editing a plan.

Remove stages. You can remove stages 2 and 3 from a multistage design.

Running an Existing Sample Plan

1. From the menus choose:
   Analyze > **Complex Samples** > **Select a Sample...**
2. Select **Draw a sample** and choose a plan file to run.
3. Click **Next** to continue through the Wizard.
4. Review the sampling plan in the Plan Summary step, and then click **Next**.
5. The individual steps containing stage information are skipped when executing a sample plan. You can now go on to the Finish step at any time.

Optionally, you can specify stages that have already been sampled.

CSPLAN and CSSELECT Commands Additional Features

The command syntax language also allows you to:

- Specify custom names for output variables.
- Control the output in the Viewer. For example, you can suppress the stagewise summary of the plan that is displayed if a sample is designed or modified, suppress the summary of the distribution of sampled cases by strata that is shown if the sample design is executed, and request a case processing summary.
- Choose a subset of variables in the active dataset to write to an external sample file or to a different dataset.

See the *Command Syntax Reference* for complete syntax information.

Chapter 3. Preparing a Complex Sample for Analysis

The Analysis Preparation Wizard guides you through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, you should have a sample drawn according to a complex design.

Creating a new plan is most useful when you do not have access to the sampling plan file used to draw the sample (recall that the sampling plan contains a default analysis plan). If you do have access to the sampling plan file used to draw the sample, you can use the default analysis plan contained in the sampling plan file or override the default analysis specifications and save your changes to a new file.

Creating a New Analysis Plan

1. From the menus choose:

   Analyze > **Complex Samples** > **Prepare for Analysis...**
2. Select **Create a plan file**, and choose a plan filename to which you will save the analysis plan.
3. Click **Next** to continue through the Wizard.
4. Specify the variable containing sample weights in the Design Variables step, optionally defining strata and clusters.
5. You can now click **Finish** to save the plan.

Optionally, in further steps you can:

- Select the method for estimating standard errors in the Estimation Method step.
- Specify the number of units sampled or the inclusion probability per unit in the Size step.
- Add a second or third stage to the design.
- Paste your selections as command syntax.

Analysis Preparation Wizard: Design Variables

This step allows you to identify the stratification and clustering variables and define sample weights. You can also provide a label for the stage.

Strata. The cross-classification of stratification variables defines distinct subpopulations, or strata. Your total sample represents the combination of independent samples from each stratum.

Clusters. Cluster variables define groups of observational units, or clusters. Samples drawn in multiple stages select clusters in the earlier stages and then subsample units from the selected clusters. When analyzing a data file obtained by sampling clusters with replacement, you should include the duplication index as a cluster variable.

Sample Weight. You must provide sample weights in the first stage. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. You can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

Note: The source variable list has the same contents across steps of the Wizard. In other words, variables removed from the source list in a particular step are removed from the list in all steps. Variables returned to the source list show up in all steps.

Tree Controls for Navigating the Analysis Wizard

At the left side of each step of the Analysis Wizard is an outline of all the steps. You can navigate the Wizard by clicking on the name of an enabled step in the outline. Steps are enabled as long as all previous steps are valid—that is, as long as each previous step has been given the minimum required specifications for that step. For more information on why a given step may be invalid, see the Help for individual steps.

Analysis Preparation Wizard: Estimation Method

This step allows you to specify an estimation method for the stage.

WR (sampling with replacement). WR estimation does not include a correction for sampling from a finite population (FPC) when estimating the variance under the complex sampling design. You can choose to include or exclude the FPC when estimating the variance under simple random sampling (SRS).

Choosing not to include the FPC for SRS variance estimation is recommended when the analysis weights have been scaled so that they do not add up to the population size. The SRS variance estimate is used in computing statistics like the design effect. WR estimation can be specified only in the final stage of a design; the Wizard will not allow you to add another stage if you select WR estimation.

Equal WOR (equal probability sampling without replacement). Equal WOR estimation includes the finite population correction and assumes that units are sampled with equal probability. Equal WOR can be specified in any stage of a design.

Unequal WOR (unequal probability sampling without replacement). In addition to using the finite population correction, Unequal WOR accounts for sampling units (usually clusters) selected with unequal probability. This estimation method is available only in the first stage.

Analysis Preparation Wizard: Size

This step is used to specify inclusion probabilities or population sizes for the current stage. Sizes can be fixed or can vary across strata. For the purpose of specifying sizes, clusters specified in previous stages can be used to define strata. Note that this step is necessary only when Equal WOR is chosen as the Estimation Method.

Units. You can specify exact population sizes or the probabilities with which units were sampled.

- **Value.** A single value is applied to all strata. If **Population Sizes** is selected as the unit metric, you should enter a non-negative integer. If **Inclusion Probabilities** is selected, you should enter a value between 0 and 1, inclusive.
- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

Define Unequal Sizes

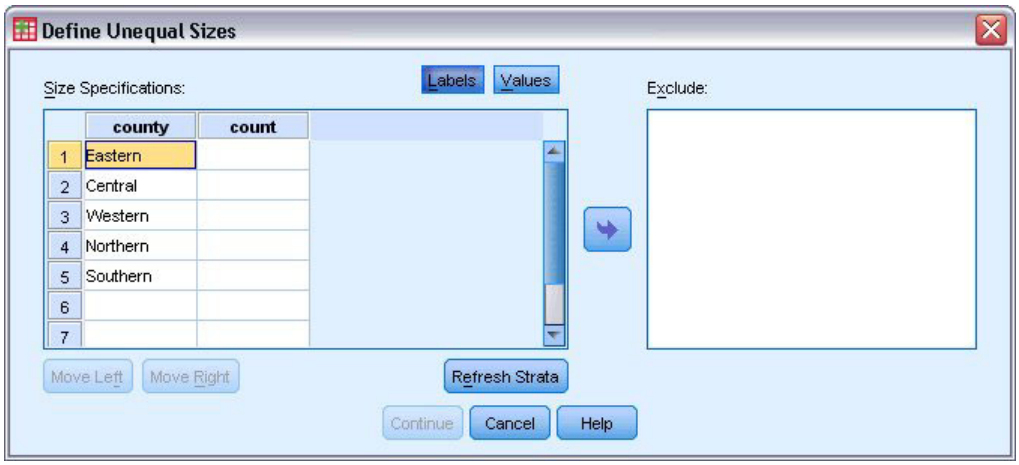

Figure 2. Define Unequal Sizes dialog box

The Define Unequal Sizes dialog box allows you to enter sizes on a per-stratum basis.

Size Specifications grid. The grid displays the cross-classifications of up to five strata or cluster variables—one stratum/cluster combination per row. Eligible grid variables include all stratification variables from the current and previous stages and all cluster variables from previous stages. Variables can be reordered within the grid or moved to the Exclude list. Enter sizes in the rightmost column. Click **Labels** or **Values** to toggle the display of value labels and data values for stratification and cluster variables in the grid cells. Cells that contain unlabeled values always show values. Click **Refresh Strata** to repopulate the grid with each combination of labeled data values for variables in the grid.

Exclude. To specify sizes for a subset of stratum/cluster combinations, move one or more variables to the Exclude list. These variables are not used to define sample sizes.

Analysis Preparation Wizard: Plan Summary

This is the last step within each stage, providing a summary of the analysis design specifications through the current stage. From here, you can either proceed to the next stage (creating it if necessary) or save the analysis specifications.

If you cannot add another stage, it is likely because:

- No cluster variable was specified in the Design Variables step.
- You selected WR estimation in the Estimation Method step.
- This is the third stage of the analysis, and the Wizard supports a maximum of three stages.

Analysis Preparation Wizard: Finish

This is the final step. You can save the plan file now or paste your selections to a syntax window.

When making changes to stages in the existing plan file, you can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If you want to save the plan to a new file, choose to **Paste the syntax generated by the Wizard into a syntax window** and change the filename in the syntax commands.

Modifying an Existing Analysis Plan

1. From the menus choose:

   Analyze > **Complex Samples** > **Prepare for Analysis...**
2. Select **Edit a plan file**, and choose a plan filename to which you will save the analysis plan.
3. Click **Next** to continue through the Wizard.
4. Review the analysis plan in the Plan Summary step, and then click **Next**.

   Subsequent steps are largely the same as for a new design. For more information, see the Help for individual steps.
5. Navigate to the Finish step, and specify a new name for the edited plan file, or choose to overwrite the existing plan file.

Optionally, you can remove stages from the plan.

Analysis Preparation Wizard: Plan Summary

This step allows you to review the analysis plan and remove stages from the plan.

Remove Stages. You can remove stages 2 and 3 from a multistage design. Since a plan must have at least one stage, you can edit but not remove stage 1 from the design.

Chapter 4. Complex Samples Plan

Complex Samples analysis procedures require analysis specifications from an analysis or sample plan file in order to provide valid results.

Plan. Specify the path of an analysis or sample plan file.

Joint Probabilities. In order to use Unequal WOR estimation for clusters drawn using a PPS WOR method, you need to specify a separate file or an open dataset containing the joint probabilities. This file or dataset is created by the Sampling Wizard during sampling.

Chapter 5. Complex Samples Frequencies

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Frequencies procedure, you can obtain univariate tabular statistics for vitamin usage among U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

Statistics. The procedure produces estimates of cell population sizes and table percentages, plus standard errors, confidence intervals, coefficients of variation, design effects, square roots of design effects, cumulative values, and unweighted counts for each estimate. Additionally, chi-square and likelihood-ratio statistics are computed for the test of equal cell proportions.

Complex Samples Frequencies Data Considerations

Data. Variables for which frequency tables are produced should be categorical. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Frequencies

1. From the menus choose:
   Analyze > **Complex Samples** > **Frequencies...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Frequencies dialog box, select at least one frequency variable.

Optionally, you can specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Frequencies Statistics

Cells. This group allows you to request estimates of the cell population sizes and table percentages.

Statistics. This group produces statistics associated with the population size or table percentage.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Cumulative values.** The cumulative estimate through each value of the variable.

Test of equal cell proportions. This produces chi-square and likelihood-ratio tests of the hypothesis that the categories of a variable have equal frequencies. Separate tests are performed for each variable.

Complex Samples Missing Values

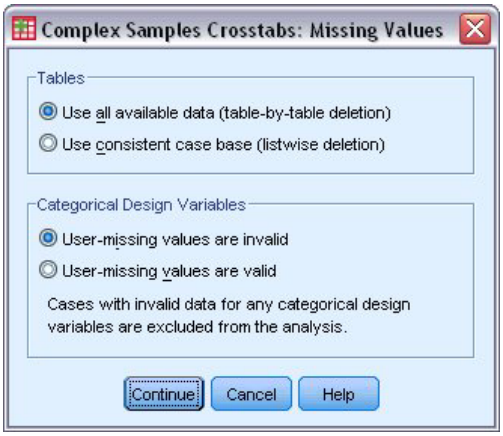

Figure 3. Missing Values dialog box

Tables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **Use consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

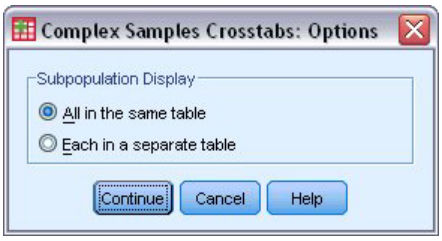

Figure 4. Options dialog box

Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter 6. Complex Samples Descriptives

The Complex Samples Descriptives procedure displays univariate summary statistics for several variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Descriptives procedure, you can obtain univariate descriptive statistics for the activity levels of U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

Statistics. The procedure produces means and sums, plus *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects for each estimate.

Complex Samples Descriptives Data Considerations

Data. Measures should be scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Descriptives

1. From the menus choose:

   Analyze > **Complex Samples** > **Descriptives...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Descriptives dialog box, select at least one measure variable.

Optionally, you can specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Descriptives Statistics

Summaries. This group allows you to request estimates of the means and sums of the measure variables. Additionally, you can request *t* tests of the estimates against a specified value.

Statistics. This group produces statistics associated with the mean or sum.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Population size.** The estimated number of units in the population.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Complex Samples Descriptives Missing Values

Statistics for Measure Variables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a variable-by-variable basis, thus the cases used to compute statistics may vary across measure variables.
- **Ensure consistent case base.** Missing values are determined across all variables, thus the cases used to compute statistics are consistent.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

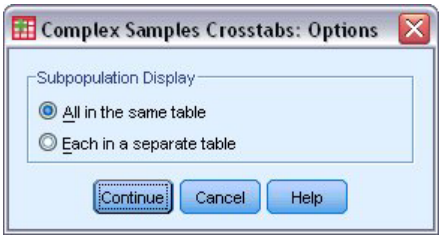

Figure 5. Options dialog box

Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter 7. Complex Samples Crosstabs

The Complex Samples Crosstabs procedure produces crosstabulation tables for pairs of selected variables and displays two-way statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Crosstabs procedure, you can obtain cross-classification statistics for smoking frequency by vitamin usage of U.S. citizens, based on the results of the National Health Interview Survey (NHIS) and with an appropriate analysis plan for this public-use data.

Statistics. The procedure produces estimates of cell population sizes and row, column, and table percentages, plus standard errors, confidence intervals, coefficients of variation, expected values, design effects, square roots of design effects, residuals, adjusted residuals, and unweighted counts for each estimate. The odds ratio, relative risk, and risk difference are computed for 2-by-2 tables. Additionally, Pearson and likelihood-ratio statistics are computed for the test of independence of the row and column variables.

Complex Samples Crosstabs Data Considerations

Data. Row and column variables should be categorical. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Crosstabs

1. From the menus choose:
   Analyze > **Complex Samples** > **Crosstabs...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Crosstabs dialog box, select at least one row variable and one column variable.

Optionally, you can specify variables to define subpopulations. Statistics are computed separately for each subpopulation.

Complex Samples Crosstabs Statistics

Cells. This group allows you to request estimates of the cell population size and row, column, and table percentages.

Statistics. This group produces statistics associated with the population size and row, column, and table percentages.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Expected values.** The expected value of the estimate, under the hypothesis of independence of the row and column variable.
- **Unweighted count.** The number of units used to compute the estimate.

- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Residuals.** The expected value is the number of cases that you would expect in the cell if there were no relationship between the two variables. A positive residual indicates that there are more cases in the cell than there would be if the row and column variables were independent.
- **Adjusted residuals.** The residual for a cell (observed minus expected value) divided by an estimate of its standard error. The resulting standardized residual is expressed in standard deviation units above or below the mean.

Summaries for 2-by-2 Tables. This group produces statistics for tables in which the row and column variable each have two categories. Each is a measure of the strength of the association between the presence of a factor and the occurrence of an event.

- **Odds ratio.** The odds ratio can be used as an estimate of relative risk when the occurrence of the factor is rare.
- **Relative risk.** The ratio of the risk of an event in the presence of the factor to the risk of the event in the absence of the factor.
- **Risk difference.** The difference between the risk of an event in the presence of the factor and the risk of the event in the absence of the factor.

Test of independence of rows and columns. This produces chi-square and likelihood-ratio tests of the hypothesis that a row and column variable are independent. Separate tests are performed for each pair of variables.

Complex Samples Missing Values

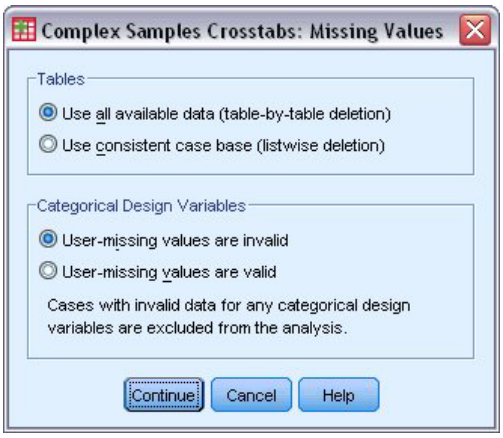

Figure 6. Missing Values dialog box

Tables. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a table-by-table basis. Thus, the cases used to compute statistics may vary across frequency or crosstabulation tables.
- **Use consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent across tables.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

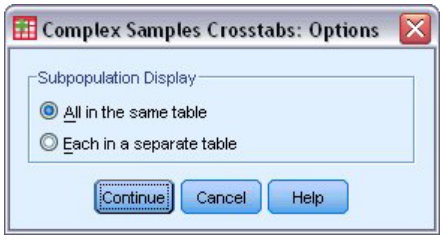

Figure 7. Options dialog box

Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter 8. Complex Samples Ratios

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, you can request statistics by subgroups, defined by one or more categorical variables.

Example. Using the Complex Samples Ratios procedure, you can obtain descriptive statistics for the ratio of current property value to last assessed value, based on the results of a statewide survey carried out according to a complex design and with an appropriate analysis plan for the data.

Statistics. The procedure produces ratio estimates, *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects.

Complex Samples Ratios Data Considerations

Data. Numerators and denominators should be positive-valued scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Ratios

1. From the menus choose:

   Analyze > **Complex Samples** > **Ratios...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Ratios dialog box, select at least one numerator variable and denominator variable.

Optionally, you can specify variables to define subgroups for which statistics are produced.

Complex Samples Ratios Statistics

Statistics. This group produces statistics associated with the ratio estimate.

- **Standard error.** The standard error of the estimate.
- **Confidence interval.** A confidence interval for the estimate, using the specified level.
- **Coefficient of variation.** The ratio of the standard error of the estimate to the estimate.
- **Unweighted count.** The number of units used to compute the estimate.
- **Population size.** The estimated number of units in the population.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

T test. You can request *t* tests of the estimates against a specified value.

Complex Samples Ratios Missing Values

Ratios. This group determines which cases are used in the analysis.

- **Use all available data.** Missing values are determined on a ratio-by-ratio basis. Thus, the cases used to compute statistics may vary across numerator-denominator pairs.
- **Ensure consistent case base.** Missing values are determined across all variables. Thus, the cases used to compute statistics are consistent.

Categorical Design Variables. This group determines whether user-missing values are valid or invalid.

Complex Samples Options

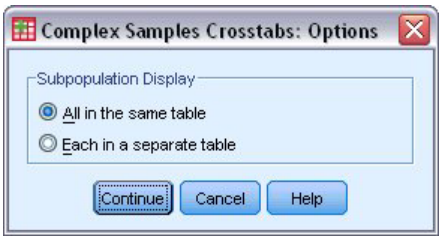

Figure 8. Options dialog box

Subpopulation Display. You can choose to have subpopulations displayed in the same table or in separate tables.

Chapter 9. Complex Samples General Linear Model

The Complex Samples General Linear Model (CSGLM) procedure performs linear regression analysis, as well as analysis of variance and covariance, for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. A grocery store chain surveyed a set of customers concerning their purchasing habits, according to a complex design. Given the survey results and how much each customer spent in the previous month, the store wants to see if the frequency with which customers shop is related to the amount they spend in a month, controlling for the gender of the customer and incorporating the sampling design.

Statistics. The procedure produces estimates, standard errors, confidence intervals, *t* tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Measures of model fit and descriptive statistics for the dependent and independent variables are also available. Additionally, you can request estimated marginal means for levels of model factors and factor interactions.

Complex Samples General Linear Model Data Considerations

Data. The dependent variable is quantitative. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining a Complex Samples General Linear Model

1. From the menus choose:
   Analyze > **Complex Samples** > **General Linear Model...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples General Linear Model dialog box, select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

Complex Samples General Linear Model

Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Intercept. The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

Complex Samples General Linear Model Statistics

Model Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Standard error.** Displays the standard error for each coefficient estimate.
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **T test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Model fit. Displays R^2 and root mean squared error statistics.

Population means of dependent variable and covariates. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- *Sequential Sidak*. This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sequential Bonferroni*. This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sidak*. This method provides tighter bounds than the Bonferroni approach.
- *Bonferroni*. This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples General Linear Model Estimated Means

The Estimated Means dialog box allows you to display the model-estimated marginal means for levels of factors and factor interactions specified in the Model subdialog box. You can also request that the overall population mean be displayed.

Term. Estimated means are computed for the selected factors and factor interactions.

Contrast. The contrast determines how hypothesis tests are set up to compare the estimated means.

- *Simple*. Compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group.
- *Deviation*. Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.
- *Difference*. Compares the mean of each level (except the first) to the mean of previous levels. They are sometimes called reverse Helmert contrasts.
- *Helmert*. Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.
- *Repeated*. Compares the mean of each level (except the last) to the mean of the subsequent level.
- *Polynomial*. Compares the linear effect, quadratic effect, cubic effect, and so on. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are often used to estimate polynomial trends.

Reference Category. The simple and deviation contrasts require a reference category or factor level against which the others are compared.

Complex Samples General Linear Model Save

Save Variables. This group allows you to save the model predicted values and residuals as new variables in the working file.

Export model as IBM SPSS Statistics data. Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- **rowtype_.** Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of freedom). There is a separate case with row type COV (or CORR) for each model parameter, plus a separate case for each of the other row types.
- **varname_.** Takes values P1, P2, ..., corresponding to an ordered list of all model parameters, for row types COV or CORR, with value labels corresponding to the parameter strings shown in the parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters, with variable labels corresponding to the parameter strings shown in the parameter estimates table, and take values according to the row type. For redundant parameters, all covariances are set to zero; correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

Note: This file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here.

Export Model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples General Linear Model Options

User-Missing Values. All design variables, as well as the dependent variable and any covariates, must have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates and estimated marginal means. Specify a value greater than or equal to 50 and less than 100.

CSGLM Command Additional Features

The command syntax language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the `CUSTOM` subcommand).
- Fix covariates at values other than their means when computing estimated marginal means (using the `EMMEANS` subcommand).
- Specify a metric for polynomial contrasts (using the `EMMEANS` subcommand).
- Specify a tolerance value for checking singularity (using the `CRITERIA` subcommand).
- Create user-specified names for saved variables (using the `SAVE` subcommand).
- Produce a general estimable function table (using the `PRINT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter 10. Complex Samples Logistic Regression

The Complex Samples Logistic Regression procedure performs logistic regression analysis on a binary or multinomial dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. A loan officer has collected past records of customers given loans at several different branches, according to a complex design. While incorporating the sample design, the officer wants to see if the probability with which a customer defaults is related to age, employment history, and amount of credit debt.

Statistics. The procedure produces estimates, exponentiated estimates, standard errors, confidence intervals, *t* tests, design effects, and square roots of design effects for model parameters, as well as the correlations and covariances between parameter estimates. Pseudo R^2 statistics, classification tables, and descriptive statistics for the dependent and independent variables are also available.

Complex Samples Logistic Regression Data Considerations

Data. The dependent variable is categorical. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Logistic Regression

1. From the menus choose:
   Analyze > **Complex Samples** > **Logistic Regression...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Logistic Regression dialog box, select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

Complex Samples Logistic Regression Reference Category

By default, the Complex Samples Logistic Regression procedure makes the highest-valued category the reference category. This dialog box allows you to specify the highest value, the lowest value, or a custom category as the reference category.

Complex Samples Logistic Regression Model

Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Intercept. The intercept is usually included in the model. If you can assume the data pass through the origin, you can exclude the intercept. Even if you include the intercept in the model, you can choose to suppress statistics related to it.

Complex Samples Logistic Regression Statistics

Model Fit. Controls the display of statistics that measure the overall model performance.

- **Pseudo R-square.** The R^2 statistic from linear regression does not have an exact counterpart among logistic regression models. There are, instead, multiple measures that attempt to mimic the properties of the R^2 statistic.
- **Classification table.** Displays the tabulated cross-classifications of the observed category by the model-predicted category on the dependent variable.

Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Exponentiated estimate.** Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **Standard error.** Displays the standard error for each coefficient estimate.
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.

- **T test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Summary statistics for model variables. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- *Sequential Sidak*. This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sequential Bonferroni*. This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sidak*. This method provides tighter bounds than the Bonferroni approach.
- *Bonferroni*. This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples Logistic Regression Odds Ratios

The Odds Ratios dialog box allows you to display the model-estimated odds ratios for specified factors and covariates. A separate set of odds ratios is computed for each category of the dependent variable except the reference category.

Factors. For each selected factor, displays the ratio of the odds at each category of the factor to the odds at the specified reference category.

Covariates. For each selected covariate, displays the ratio of the odds at the covariate's mean value plus the specified units of change to the odds at the mean.

When computing odds ratios for a factor or covariate, the procedure fixes all other factors at their highest levels and all other covariates at their means. If a factor or covariate interacts with other predictors in the model, then the odds ratios depend not only on the change in the specified variable but also on the values of the variables with which it interacts. If a specified covariate interacts with itself in the model (for example, *age*age*), then the odds ratios depend on both the change in the covariate and the value of the covariate.

Complex Samples Logistic Regression Save

Save Variables. This group allows you to save the model-predicted category and predicted probabilities as new variables in the active dataset.

Export model as IBM SPSS Statistics data. Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- **rowtype_.** Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of freedom). There is a separate case with row type COV (or CORR) for each model parameter, plus a separate case for each of the other row types.
- **varname_.** Takes values P1, P2, ..., corresponding to an ordered list of all model parameters, for row types COV or CORR, with value labels corresponding to the parameter strings shown in the parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters, with variable labels corresponding to the parameter strings shown in the parameter estimates table, and take values according to the row type. For redundant parameters, all covariances are set to zero; correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

Note: This file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here.

Export Model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples Logistic Regression Options

Estimation. This group gives you control of various criteria used in the model estimation.

- **Maximum Iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum Step-Halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Limit iterations based on change in parameter estimates.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be non-negative.
- **Limit iterations based on change in log-likelihood.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be non-negative.
- **Check for complete separation of data points.** When selected, the algorithm performs tests to ensure that the parameter estimates have unique values. Separation occurs when the procedure can produce a model that correctly classifies every case.

- **Display iteration history.** Displays parameter estimates and statistics at every *n* iterations beginning with the 0^{th} iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of *n*.

User-Missing Values. All design variables, as well as the dependent variable and any covariates, must have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates, exponentiated coefficient estimates, and odds ratios. Specify a value greater than or equal to 50 and less than 100.

CSLOGISTIC Command Additional Features

The command syntax language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the `CUSTOM` subcommand).
- Fix values of other model variables when computing odds ratios for factors and covariates (using the `ODDSRATIOS` subcommand).
- Specify a tolerance value for checking singularity (using the `CRITERIA` subcommand).
- Create user-specified names for saved variables (using the `SAVE` subcommand).
- Produce a general estimable function table (using the `PRINT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter 11. Complex Samples Ordinal Regression

The Complex Samples Ordinal Regression procedure performs regression analysis on a binary or ordinal dependent variable for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Example. Representatives considering a bill before the legislature are interested in whether there is public support for the bill and how support for the bill is related to voter demographics. Pollsters design and conduct interviews according to a complex sampling design. Using Complex Samples Ordinal Regression, you can fit a model for the level of support for the bill based upon voter demographics.

Complex Samples Ordinal Regression Data Considerations

Data. The dependent variable is ordinal. Factors are categorical. Covariates are quantitative variables that are related to the dependent variable. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Obtaining Complex Samples Ordinal Regression

1. From the menus choose:

   Analyze > **Complex Samples** > **Ordinal Regression...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. In the Complex Samples Ordinal Regression dialog box, select a dependent variable.

Optionally, you can:

- Select variables for factors and covariates, as appropriate for your data.
- Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable, although variances are still properly estimated based on the entire dataset.
- Select a link function.

Link function. The link function is a transformation of the cumulative probabilities that allows estimation of the model. The following five link functions are available.

- **Logit.** $f(x)=\log(x/(1-x))$. Typically used for evenly distributed categories.
- **Complementary log-log.** $f(x)=\log(-\log(1-x))$. Typically used when higher categories are more probable.
- **Negative log-log.** $f(x)=-\log(-\log(x))$. Typically used when lower categories are more probable.
- **Probit.** $f(x)=\Phi^{-1}(x)$. Typically used when the latent variable is normally distributed.
- **Cauchit (inverse Cauchy).** $f(x)=\tan(\pi(x-0.5))$. Typically used when the latent variable has many extreme values.

Complex Samples Ordinal Regression Response Probabilities

The Response Probabilities dialog box allows you to specify whether the cumulative probability of a response (that is, the probability of belonging up to and including a particular category of the dependent variable) increases with increasing or decreasing values of the dependent variable.

Complex Samples Ordinal Regression Model

Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Complex Samples Ordinal Regression Statistics

Model Fit. Controls the display of statistics that measure the overall model performance.

- **Pseudo R-square.** The R^2 statistic from linear regression does not have an exact counterpart among ordinal regression models. There are, instead, multiple measures that attempt to mimic the properties of the R^2 statistic.
- **Classification table.** Displays the tabulated cross-classifications of the observed category by the model-predicted category on the dependent variable.

Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Exponentiated estimate.** Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **Standard error.** Displays the standard error for each coefficient estimate.

- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **T test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.
- **Square root of design effect.** This is a measure, expressed in units comparable to those of the standard error, of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Parallel Lines. This group allows you to request statistics associated with a model with nonparallel lines where a separate regression line is fitted for each response category (except the last).

- **Wald test.** Produces a test of the null hypothesis that regression parameters are equal for all cumulative responses. The model with nonparallel lines is estimated and the Wald test of equal parameters is applied.
- **Parameter estimates.** Displays estimates of the coefficients and standard errors for the model with nonparallel lines.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the coefficients of the model with nonparallel lines.

Summary statistics for model variables. Displays summary information about the dependent variable, covariates, and factors.

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Complex Samples Hypothesis Tests

Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- *Sequential Sidak*. This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sequential Bonferroni*. This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sidak*. This method provides tighter bounds than the Bonferroni approach.

- *Bonferroni*. This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Complex Samples Ordinal Regression Odds Ratios

The Odds Ratios dialog box allows you to display the model-estimated cumulative odds ratios for specified factors and covariates. This feature is only available for models using the Logit link function. A single cumulative odds ratio is computed for all categories of the dependent variable except the last; the proportional odds model postulates that they are all equal.

Factors. For each selected factor, displays the ratio of the cumulative odds at each category of the factor to the odds at the specified reference category.

Covariates. For each selected covariate, displays the ratio of the cumulative odds at the covariate's mean value plus the specified units of change to the odds at the mean.

When computing odds ratios for a factor or covariate, the procedure fixes all other factors at their highest levels and all other covariates at their means. If a factor or covariate interacts with other predictors in the model, then the odds ratios depend not only on the change in the specified variable but also on the values of the variables with which it interacts. If a specified covariate interacts with itself in the model (for example, *age*age*), then the odds ratios depend on both the change in the covariate and the value of the covariate.

Complex Samples Ordinal Regression Save

Save Variables. This group allows you to save the model-predicted category, probability of predicted category, probability of observed category, cumulative probabilities, and predicted probabilities as new variables in the active dataset.

Export model as IBM SPSS Statistics data. Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- **rowtype_.** Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of freedom). There is a separate case with row type COV (or CORR) for each model parameter, plus a separate case for each of the other row types.
- **varname_.** Takes values P1, P2, ..., corresponding to an ordered list of all model parameters, for row types COV or CORR, with value labels corresponding to the parameter strings shown in the parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters, with variable labels corresponding to the parameter strings shown in the parameter estimates table, and take values according to the row type. For redundant parameters, all covariances are set to zero; correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

Note: This file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here.

Export model as XML. Saves the parameter estimates and the parameter covariance matrix, if selected, in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Complex Samples Ordinal Regression Options

Estimation Method. You can select a parameter estimation method; choose between Newton-Raphson, Fisher scoring, or a hybrid method in which Fisher scoring iterations are performed before switching to the Newton-Raphson method. If convergence is achieved during the Fisher scoring phase of the hybrid method before the maximum number of Fisher iterations is reached, the algorithm continues with the Newton-Raphson method.

Estimation. This group gives you control of various criteria used in the model estimation.

- **Maximum Iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum Step-Halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Limit iterations based on change in parameter estimates.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be non-negative.
- **Limit iterations based on change in log-likelihood.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be non-negative.
- **Check for complete separation of data points.** When selected, the algorithm performs tests to ensure that the parameter estimates have unique values. Separation occurs when the procedure can produce a model that correctly classifies every case.
- **Display iteration history.** Displays parameter estimates and statistics at every *n* iterations beginning with the 0^{th} iteration (the initial estimates). If you choose to print the iteration history, the last iteration is always printed regardless of the value of *n*.

User-Missing Values. Scale design variables, as well as the dependent variable and any covariates, should have valid data. Cases with invalid data for any of these variables are deleted from the analysis. These controls allow you to decide whether user-missing values are treated as valid among the strata, cluster, subpopulation, and factor variables.

Confidence Interval. This is the confidence interval level for coefficient estimates, exponentiated coefficient estimates, and odds ratios. Specify a value greater than or equal to 50 and less than 100.

CSORDINAL Command Additional Features

The command syntax language also allows you to:

- Specify custom tests of effects versus a linear combination of effects or a value (using the `CUSTOM` subcommand).
- Fix values of other model variables at values other than their means when computing cumulative odds ratios for factors and covariates (using the `ODDSRATIOS` subcommand).
- Use unlabeled values as custom reference categories for factors when odds ratios are requested (using the `ODDSRATIOS` subcommand).
- Specify a tolerance value for checking singularity (using the `CRITERIA` subcommand).
- Produce a general estimable function table (using the `PRINT` subcommand).
- Save more than 25 probability variables (using the `SAVE` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter 12. Complex Samples Cox Regression

The Complex Samples Cox Regression procedure performs survival analysis for samples drawn by complex sampling methods. Optionally, you can request analyses for a subpopulation.

Examples. A government law enforcement agency is concerned about recidivism rates in their area of jurisdiction. One of the measures of recidivism is the time until second arrest for offenders. The agency would like to model time to rearrest using Cox Regression but are worried the proportional hazards assumption is invalid across age categories.

Medical researchers are investigating survival times for patients exiting a rehabilitation program post-ischemic stroke. There is the potential for multiple cases per subject, since patient histories change as the occurrence of significant nondeath events are noted and the times of these events recorded. The sample is also left-truncated in the sense that the observed survival times are "inflated" by the length of rehabilitation, because while the onset of risk starts at the time of the ischemic stroke, only patients who survive past the rehabilitation program are in the sample.

Complex Samples Cox Regression Data Considerations

Survival Time. The procedure applies Cox regression to analysis of survival times—that is, the length of time before the occurrence of an event. There are two ways to specify the survival time, depending upon the start time of the interval:

- **Time=0.** Commonly, you will have complete information on the start of the interval for each subject and will simply have a variable containing end times (or create a single variable with end times from Date & Time variables; see below).
- **Varies by subject.** This is appropriate when you have **left-truncation**, also called **delayed entry**; for example, if you are analyzing survival times for patients exiting a rehabilitation program post-stroke, you might consider that their onset of risk starts at the time of the stroke. However, if your sample only includes patients who have survived the rehabilitation program, then your sample is left-truncated in the sense that the observed survival times are "inflated" by the length of rehabilitation. You can account for this by specifying the time at which they exited rehabilitation as the time of entry into the study.

Date & Time Variables. Date & Time variables cannot be used to directly define the start and end of the interval; if you have Date & Time variables, you should use them to create variables containing survival times. If there is no left-truncation, simply create a variable containing end times based upon the difference between the date of entry into the study and the observation date. If there is left-truncation, create a variable containing start times, based upon the difference between the date of the start of the study and the date of entry, and a variable containing end times, based upon the difference between the date of the start of the study and the date of observation.

Event Status. You need a variable that records whether the subject experienced the event of interest within the interval. Subjects for whom the event has not occurred are right-censored.

Subject Identifier. You can easily incorporate piecewise-constant, time-dependent predictors by splitting the observations for a single subject across multiple cases. For example, if you are analyzing survival times for patients post-stroke, variables representing their medical history should be useful as predictors. Over time, they may experience major medical events that alter their medical history. The following table shows how to structure such a dataset: *Patient ID* is the subject identifier, *End time* defines the observed intervals, *Status* records major medical events, and *Prior history of heart attack* and *Prior history of hemorrhaging* are piecewise-constant, time-dependent predictors.

Table 1. Data structure for incorporating piecewise-constant time-dependent predictors.

Patient ID	*End time*	*Status*	*Prior history of heart attack*	*Prior history of hemorrhaging*
1	5	Heart Attack	No	No
1	7	Hemorrhaging	Yes	No
1	8	Died	Yes	Yes
2	24	Died	No	No
3	8	Heart Attack	No	No
3	15	Died	Yes	No

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Typically, Cox regression models assume proportional hazards—that is, the ratio of hazards from one case to another should not vary over time. If this assumption does not hold, you may need to add time-dependent predictors to the model.

Kaplan-Meier Analysis. If you do not select any predictors (or do not enter any selected predictors into the model) and choose the product limit method for computing the baseline survival curve on the Options tab, the procedure performs a Kaplan-Meier type of survival analysis.

To Obtain Complex Samples Cox Regression

1. From the menus choose:
   Analyze > **Complex Samples** > **Cox Regression...**
2. Select a plan file. Optionally, select a custom joint probabilities file.
3. Click **Continue**.
4. Specify the survival time by selecting the entry and exit times from the study.
5. Select an event status variable.
6. Click Define Event and define at least one event value.

Optionally, you can select a subject identifier.

Define Event

Specify the values that indicate a terminal event has occurred.

- **Individual value(s).** Specify one or more values by entering them into the grid or selecting them from a list of values with defined value labels.
- **Range of values.** Specify a range of values by entering the minimum and maximum values or selecting values from a list with defined value labels.

Predictors

The Predictors tab allows you to specify the factors and covariates used to build model effects.

Factors. Factors are categorical predictors; they can be numeric or string.

Covariates. Covariates are scale predictors; they must be numeric.

Time-Dependent Predictors. There are certain situations in which the proportional hazards assumption does not hold. That is, hazard ratios change across time; the values of one (or more) of your predictors

are different at different time points. In such cases, you need to specify time-dependent predictors. See the topic "Define Time-Dependent Predictor" for more information. Time-dependent predictors can be selected as factors or covariates.

Define Time-Dependent Predictor

The Define Time-Dependent Predictor dialog box allows you to create a predictor that is dependent upon the built-in time variable, *T_*. You can use this variable to define time-dependent covariates in two general ways:

- If you want to estimate an extended Cox regression model that allows nonproportional hazards, you can do so by defining your time-dependent predictor as a function of the time variable *T_* and the covariate in question. A common example would be the simple product of the time variable and the predictor, but more complex functions can be specified as well.
- Some variables may have different values at different time periods but aren't systematically related to time. In such cases, you need to define a **segmented time-dependent predictor**, which can be done using logical expressions. Logical expressions take the value 1 if true and 0 if false. Using a series of logical expressions, you can create your time-dependent predictor from a set of measurements. For example, if you have blood pressure measured once a week for the four weeks of your study (identified as *BP1* to *BP4*), you can define your time-dependent predictor as (*T_* < 1) * *BP1* + (*T_* >= 1 & *T_* < 2) * *BP2* + (*T_* >= 2 & *T_* < 3) * *BP3* + (*T_* >= 3 & *T_* < 4) * *BP4*. Notice that exactly one of the terms in parentheses will be equal to 1 for any given case and the rest will all equal 0. In other words, this function means that if time is less than one week, use *BP1*; if it is more than one week but less than two weeks, use *BP2*; and so on.

  Note: If your segmented, time-dependent predictor is constant within segments, as in the blood pressure example given above, it may be easier for you to specify the piecewise-constant, time-dependent predictor by splitting subjects across multiple cases. See the discussion on Subject Identifiers in Chapter 12, "Complex Samples Cox Regression," on page 41 for more information.

In the Define Time-Dependent Predictor dialog box, you can use the function-building controls to build the expression for the time-dependent covariate, or you can enter it directly in the Numeric Expression text area. Note that string constants must be enclosed in quotation marks or apostrophes, and numeric constants must be typed in American format, with the dot as the decimal delimiter. The resulting variable is given the name you specify and should be included as a factor or covariate on the Predictors tab.

Subgroups

Baseline Strata. A separate baseline hazard and survival function is computed for each value of this variable, while a single set of model coefficients is estimated across strata.

Subpopulation Variable. Specify a variable to define a subpopulation. The analysis is performed only for the selected category of the subpopulation variable.

Model

Specify Model Effects. By default, the procedure builds a main-effects model using the factors and covariates specified in the main dialog box. Alternatively, you can build a custom model that includes interaction effects and nested terms.

Non-Nested Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term for all selected variables.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Nested Terms

You can build nested terms for your model in this procedure. Nested terms are useful for modeling the effect of a factor or covariate whose values do not interact with the levels of another factor. For example, a grocery store chain may follow the spending habits of its customers at several store locations. Since each customer frequents only one of these locations, the *Customer* effect can be said to be **nested within** the *Store location* effect.

Additionally, you can include interaction effects, such as polynomial terms involving the same covariate, or add multiple levels of nesting to the nested term.

Limitations. Nested terms have the following restrictions:

- All factors within an interaction must be unique. Thus, if *A* is a factor, then specifying *A*A* is invalid.
- All factors within a nested effect must be unique. Thus, if *A* is a factor, then specifying *A(A)* is invalid.
- No effect can be nested within a covariate. Thus, if *A* is a factor and *X* is a covariate, then specifying *A(X)* is invalid.

Statistics

Sample design information. Displays summary information about the sample, including the unweighted count and the population size.

Event and censoring summary. Displays summary information about the number and percentage of censored cases.

Risk set at event times. Displays number of events and number at risk for each event time in each baseline stratum.

Parameters. This group allows you to control the display of statistics related to the model parameters.

- **Estimate.** Displays estimates of the coefficients.
- **Exponentiated estimate.** Displays the base of the natural logarithm raised to the power of the estimates of the coefficients. While the estimate has nice properties for statistical testing, the exponentiated estimate, or exp(B), is easier to interpret.
- **Standard error.** Displays the standard error for each coefficient estimate.
- **Confidence interval.** Displays a confidence interval for each coefficient estimate. The confidence level for the interval is set in the Options dialog box.
- **t-test.** Displays a *t* test of each coefficient estimate. The null hypothesis for each test is that the value of the coefficient is 0.
- **Covariances of parameter estimates.** Displays an estimate of the covariance matrix for the model coefficients.
- **Correlations of parameter estimates.** Displays an estimate of the correlation matrix for the model coefficients.
- **Design effect.** The ratio of the variance of the estimate to the variance obtained by assuming that the sample is a simple random sample. This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

- **Square root of design effect.** This is a measure of the effect of specifying a complex design, where values further from 1 indicate greater effects.

Model Assumptions. This group allows you to produce a test of the proportional hazards assumption. The test compares the fitted model to an alternative model that includes time-dependent predictors *x**_TF* for each predictor *x*, where *_TF* is the specified time function.

- **Time Function.** Specifies the form of *_TF* for the alternative model. For the **identity** function, *_TF=T_*. For the **log** function, *_TF*=log(*T_*). For **Kaplan-Meier**, *_TF*=1−$S_{KM}(T\_)$, where $S_{KM}(.)$ is the Kaplan-Meier estimate of the survival function. For **rank**, *_TF* is the rank-order of *T_* among the observed end times.
- **Parameter estimates for alternative model.** Displays the estimate, standard error, and confidence interval for each parameter in the alternative model.
- **Covariance matrix for alternative model.** Displays the matrix of estimated covariances between parameters in the alternative model.

Baseline survival and cumulative hazard functions. Displays the baseline survival function and baseline cumulative hazards function along with their standard errors.

Note: If time-dependent predictors defined on the Predictors tab are included in the model, this option is not available.

Plots

The Plots tab allows you to request plots of the hazard function, survival function, log-minus-log of the survival function, and one minus the survival function. You can also choose to plot confidence intervals along the specified functions; the confidence level is set on the Options tab.

Predictor patterns. You can specify a pattern of predictor values to be used for the requested plots and the exported survival file on the Export tab. Note that these options are not available if time-dependent predictors defined on the Predictors tab are included in the model.

- **Plot Factors at.** By default, each factor is evaluated at its highest level. Enter or select a different level if wanted. Alternatively, you can choose to plot separate lines for each level of a single factor by selecting the check box for that factor.
- **Plot Covariates at.** Each covariate is evaluated at its mean. Enter or select a different value if you want.

Hypothesis Tests

Test Statistic. This group allows you to select the type of statistic used for testing hypotheses. You can choose between *F*, adjusted *F*, chi-square, and adjusted chi-square.

Sampling Degrees of Freedom. This group gives you control over the sampling design degrees of freedom used to compute *p* values for all test statistics. If based on the sampling design, the value is the difference between the number of primary sampling units and the number of strata in the first stage of sampling. Alternatively, you can set a custom degrees of freedom by specifying a positive integer.

Adjustment for Multiple Comparisons. When performing hypothesis tests with multiple contrasts, the overall significance level can be adjusted from the significance levels for the included contrasts. This group allows you to choose the adjustment method.

- **Least significant difference.** This method does not control the overall probability of rejecting the hypotheses that some linear contrasts are different from the null hypothesis values.
- *Sequential Sidak*. This is a sequentially step-down rejective Sidak procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.

- *Sequential Bonferroni*. This is a sequentially step-down rejective Bonferroni procedure that is much less conservative in terms of rejecting individual hypotheses but maintains the same overall significance level.
- *Sidak*. This method provides tighter bounds than the Bonferroni approach.
- *Bonferroni*. This method adjusts the observed significance level for the fact that multiple contrasts are being tested.

Save

Save Variables. This group allows you to save model-related variables to the active dataset for further use in diagnostics and reporting of results. Note that none of these are available when time-dependent predictors are included in the model.

- **Survival function.** Saves the probability of survival (the value of the survival function) at the observed time and predictor values for each case.
- **Lower bound of confidence interval for survival function.** Saves the lower bound of the confidence interval for the survival function at the observed time and predictor values for each case.
- **Upper bound of confidence interval for survival function.** Saves the upper bound of the confidence interval for the survival function at the observed time and predictor values for each case.
- **Cumulative hazard function.** Saves the cumulative hazard, or –ln(survival), at the observed time and predictor values for each case.
- **Lower bound of confidence interval for cumulative hazard function.** Saves the lower bound of the confidence interval for the cumulative hazard function at the observed time and predictor values for each case.
- **Upper bound of confidence interval for cumulative hazard function.** Saves the upper bound of the confidence interval for the cumulative hazard function at the observed time and predictor values for each case.
- **Predicted value of linear predictor.** Saves the linear combination of reference value corrected predictors times regression coefficients. The linear predictor is the ratio of the hazard function to the baseline hazard. Under the proportional hazards model, this value is constant across time.
- **Schoenfeld residual.** For each uncensored case and each nonredundant parameter in the model, the Schoenfeld residual is the difference between the observed value of the predictor associated with the model parameter and the expected value of the predictor for cases in the risk set at the observed event time. Schoenfeld residuals can be used to help assess the proportional hazards assumption; for example, for a predictor x, plots of the Schoenfeld residuals for the time-dependent predictor x*ln($T\_$) versus time should show a horizontal line at 0 if proportional hazards holds. A separate variable is saved for each nonredundant parameter in the model. Schoenfeld residuals are only computed for uncensored cases.
- **Martingale residual.** For each case, the martingale residual is the difference between the observed censoring (0 if censored, 1 if not) and the expectation of an event during the observation time.
- **Deviance residual.** Deviance residuals are martingale residuals "adjusted" to appear more symmetrical about 0. Plots of deviance residuals against predictors should reveal no patterns.
- **Cox-Snell residual.** For each case, the Cox-Snell residual is the expectation of an event during the observation time, or the observed censoring minus the martingale residual.
- **Score residual.** For each case and each nonredundant parameter in the model, the score residual is the contribution of the case to the first derivative of the pseudo-likelihood. A separate variable is saved for each nonredundant parameter in the model.
- **DFBeta residual.** For each case and each nonredundant parameter in the model, the DFBeta residual approximates the change in the value of the parameter estimate when the case is removed from the model. Cases with relatively large DFBeta residuals may be exerting undue influence on the analysis. A separate variable is saved for each nonredundant parameter in the model.
- **Aggregated residuals.** When multiple cases represent a single subject, the aggregated residual for a subject is simply the sum of the corresponding case residuals over all cases belonging to the same

subject. For Schoenfeld's residual, the aggregated version is the same as that of the non-aggregated version because Schoenfeld's residual is only defined for uncensored cases. These residuals are only available when a subject identifier is specified on the Time and Event tab.

Names of Saved Variables. Automatic name generation ensures that you keep all your work. Custom names allow you to discard/replace results from previous runs without first deleting the saved variables in the Data Editor.

Export

Export model as IBM SPSS Statistics data. Writes a dataset in IBM SPSS Statistics format containing the parameter correlation or covariance matrix with parameter estimates, standard errors, significance values, and degrees of freedom. The order of variables in the matrix file is as follows.

- **rowtype_.** Takes values (and value labels), COV (Covariances), CORR (Correlations), EST (Parameter estimates), SE (Standard errors), SIG (Significance levels), and DF (Sampling design degrees of freedom). There is a separate case with row type COV (or CORR) for each model parameter, plus a separate case for each of the other row types.
- **varname_.** Takes values P1, P2, ..., corresponding to an ordered list of all model parameters, for row types COV or CORR, with value labels corresponding to the parameter strings shown in the parameter estimates table. The cells are blank for other row types.
- **P1, P2, ...** These variables correspond to an ordered list of all model parameters, with variable labels corresponding to the parameter strings shown in the parameter estimates table, and take values according to the row type. For redundant parameters, all covariances are set to zero; correlations are set to the system-missing value; all parameter estimates are set at zero; and all standard errors, significance levels, and residual degrees of freedom are set to the system-missing value.

Note: This file is not immediately usable for further analyses in other procedures that read a matrix file unless those procedures accept all the row types exported here.

Export survival function as IBM SPSS Statistics data. Writes a dataset in IBM SPSS Statistics format containing the survival function; standard error of the survival function; upper and lower bounds of the confidence interval of the survival function; and the cumulative hazards function for each failure or event time, evaluated at the baseline and at the predictor patterns specified on the Plot tab. The order of variables in the matrix file is as follows.

- **Baseline strata variable.** Separate survival tables are produced for each value of the strata variable.
- **Survival time variable.** The event time; a separate case is created for each unique event time.
- **Sur_0, LCL_Sur_0, UCL_Sur_0.** Baseline survival function and the upper and lower bounds of its confidence interval.
- **Sur_R, LCL_Sur_R, UCL_Sur_R.** Survival function evaluated at the "reference" pattern (see the pattern values table in the output) and the upper and lower bounds of its confidence interval.
- **Sur_#.#, LCL_Sur_#.#, UCL_Sur_#.#, ...** Survival function evaluated at each of the predictor patterns specified on the Plots tab and the upper and lower bounds of their confidence intervals. See the pattern values table in the output to match patterns with the number #.#.
- **Haz_0, LCL_Haz_0, UCL_Haz_0.** Baseline cumulative hazard function and the upper and lower bounds of its confidence interval.
- **Haz_R, LCL_Haz_R, UCL_Haz_R.** Cumulative hazard function evaluated at the "reference" pattern (see the pattern values table in the output) and the upper and lower bounds of its confidence interval.
- **Haz_#.#, LCL_Haz_#.#, UCL_Haz_#.#, ...** Cumulative hazard function evaluated at each of the predictor patterns specified on the Plots tab and the upper and lower bounds of their confidence intervals. See the pattern values table in the output to match patterns with the number #.#.

Export model as XML. Saves all information needed to predict the survival function, including parameter estimates and the baseline survival function, in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Options

Estimation. These controls specify criteria for estimation of regression coefficients.

- **Maximum Iterations.** The maximum number of iterations the algorithm will execute. Specify a non-negative integer.
- **Maximum Step-Halving.** At each iteration, the step size is reduced by a factor of 0.5 until the log-likelihood increases or maximum step-halving is reached. Specify a positive integer.
- **Limit iterations based on change in parameter estimates.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the parameter estimates is less than the value specified, which must be positive.
- **Limit iterations based on change in log-likelihood.** When selected, the algorithm stops after an iteration in which the absolute or relative change in the log-likelihood function is less than the value specified, which must be positive.
- **Display iteration history.** Displays the iteration history for the parameter estimates and pseudo log-likelihood and prints the last evaluation of the change in parameter estimates and pseudo log-likelihood. The iteration history table prints every *n* iterations beginning with the *0*th iteration (the initial estimates), where *n* is the value of the increment. If the iteration history is requested, then the last iteration is always displayed regardless of *n*.
- **Tie breaking method for parameter estimation.** When there are tied observed failure times, one of these methods is used to break the ties. The Efron method is more computationally expensive.

Survival Functions. These controls specify criteria for computations involving the survival function.

- **Method for estimating baseline survival functions.** The **Breslow** (or Nelson-Aalan or empirical) method estimates the baseline cumulative hazard by a nondecreasing step function with steps at the observed failure times, then computes the baseline survival by the relation survival=exp(−cumulative hazard). The **Efron** method is more computationally expensive and reduces to the Breslow method when there are no ties. The **product limit** method estimates the baseline survival by a non-increasing right continuous function; when there are no predictors in the model, this method reduces to Kaplan-Meier estimation.
- **Confidence intervals of survival functions.** The confidence interval can be calculated in three ways: in original units, via a log transformation, or a log-minus-log transformation. Only the log-minus-log transformation guarantees that the bounds of the confidence interval will lie between 0 and 1, but the log transformation generally seems to perform "best."

User Missing Values. All variables must have valid values for a case to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical models (including factors, event, strata, and subpopulation variables) and sampling design variables.

Confidence interval(%). This is the confidence interval level used for coefficient estimates, exponentiated coefficient estimates, survival function estimates, and cumulative hazard function estimates. Specify a value greater than or equal to 0, and less than 100.

CSCOXREG Command Additional Features

The command language also allows you to:

- Perform custom hypothesis tests (using the `CUSTOM` subcommand and `/PRINT LMATRIX`).
- Tolerance specification (using `/CRITERIA SINGULAR`).
- General estimable function table (using `/PRINT GEF`).
- Multiple predictor patterns (using multiple `PATTERN` subcommands).

- Maximum number of saved variables when a rootname is specified (using the SAVE subcommand). The dialog honors the CSCOXREG default of 25 variables.

See the *Command Syntax Reference* for complete syntax information.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

T

U

Printed in USA